TITLE:

Integrated Parallel Data Extraction from Comparable Corpora for Statistical Machine Translation(Abstract_要旨)

AUTHOR(S):

Chu, Chenhui

CITATION:

Chu, Chenhui. Integrated Parallel Data Extraction from Comparable Corpora for Statistical Machine Translation. 京都大学, 2015, 博士(情報学)

ISSUE DATE:

2015-03-23

URL:

https://doi.org/10.14989/doctor.k19107

RIGHT:

京都大学学術情報リポジトリ
KURENAI 紅
Kyoto University Research Information Repository

（ 続紙 1 ）

京都大学	博士（情報学）	氏名	褚　晨翚　Chenhui CHU
論文題目	Integrated Parallel Data Extraction from Comparable Corpora for Statistical Machine Translation （統計的機械翻訳におけるコンパラブルコーパスからの対訳データの統合的抽出）		

（論文内容の要旨）

Machine translation (MT), as a high level application of natural language processing, is a powerful tool to improve the efficiency and reduce the cost of translation. Over the last decade or two, statistical machine translation (SMT) has been the main approach in both the research community and the commercial sector. In SMT, translation knowledge is automatically acquired from parallel corpora (sentence-aligned bilingual texts), making the rapid development of MT systems for different language pairs and domains possible once parallel corpora are available. Because of the high dependence on parallel corpora, the quality and quantity of parallel corpora are crucial for SMT. However, except for a few language pairs and some specialized domains, high quality parallel corpora of sufficient size remain a scarce resource. This scarceness of parallel corpora has become the main bottleneck for SMT.

Comparable corpora are a set of monolingual corpora that describe roughly the same topic in different languages, but are not exact translation equivalents of each other. Exploiting comparable corpora for SMT is the key to addressing the scarceness of parallel corpora. The reason for this is that comparable corpora are far more available than parallel corpora, and there is a large amount of parallel data contained in the comparable texts. The main focus of this thesis is extracting the parallel data from comparable corpora to improve SMT. There are three types of parallel data in comparable corpora: bilingual lexicons, parallel sentences and parallel fragments. In this thesis, we propose novel approaches to extract these three types of parallel data from comparable corpora in an integrated framework. In addition, we exploit linguistic knowledge of common Chinese characters for Chinese-Japanese parallel data extraction as a case study. Bilingual lexicon extraction (BLE) is used for parallel sentence extraction and improving SMT accuracy. The extracted parallel sentences and fragments are used as training data for SMT. Experiments show the effectiveness of our proposed approaches for the scarceness of parallel corpora that SMT suffers from.

In Chapter 2, we propose a method for automatically constructing a more complete resource of common Chinese characters for the Chinese-Japanese language pair using freely available resources. In addition, we propose an approach exploiting common Chinese characters in Chinese word segmentation for SMT. Common Chinese characters are used for parallel sentence (Chapter 4) and fragment extraction (Chapter 5). The optimized Chinese word segmenter is used throughout this thesis work.

In Chapter 3, we present an iterative BLE system that is based on a novel combination of topic

model and context based methods, which are the two main categories of methods that have been proposed for BLE from comparable corpora in the literature. Our system does not rely on any prior knowledge and the performance can be iteratively improved. Experiments conducted on Chinese-English, Japanese-English and Chinese-Japanese Wikipedia data show the effectiveness of our proposed method.

In Chapter 4, we present a robust parallel sentence extraction system for constructing a Chinese-Japanese parallel corpus from Wikipedia. The system mainly consists of a parallel sentence candidate filter and a classifier for parallel sentence identification. We improve the system by using common Chinese characters for filtering and three novel feature sets for classification. We further apply the bilingual lexicons extracted in Chapter 3 for parallel sentence extraction. Experiments show that our system performs significantly better than the previous studies for both accuracy in parallel sentence extraction and SMT performance.

In Chapter 5, an accurate parallel fragment extraction system is proposed. In many types of comparable corpora, there are parallel fragments existing in comparable sentences that are also helpful for SMT. We propose a system that uses a word alignment to locate the parallel fragment candidates, and uses an accurate lexicon-based filter to identify the truly parallel ones. We further use common Chinese characters for the lexicon-based filter to improve its coverage. Experiments conducted on Chinese-Japanese comparable corpora indicate that our system can accurately extract parallel fragments. In addition, we show that parallel sentences and fragments can be extracted in an integrated manner from comparable corpora.

In Chapter 6, BLE together with paraphrases is proposed for the accuracy problem of SMT. The translation pairs and their feature scores in the translation model of SMT can be inaccurate, because of the quality of the unsupervised methods used for translation model learning. Estimating comparable features from comparable corpora with BLE has been proposed for the accuracy problem of SMT. However, BLE suffers from the data sparseness, which makes the comparable features inaccurate. We propose using paraphrases to address this issue. Paraphrases are used to smooth the vectors used in comparable feature estimation with BLE. Experiments conducted on Chinese-English SMT show the effectiveness of our proposed method.

In Chapter 7, we provide concluding remarks and summaries of this thesis, and outline the possible directions for future work.

注)論文内容の要旨と論文審査の結果の要旨は１頁を３８字×３６行で作成し，合わせて，３，０００字を標準とすること．
論文内容の要旨を英語で記入する場合は，４００～１，１００wordsで作成し
審査結果の要旨は日本語５００～２，０００字程度で作成すること．

（続紙 ２ ）

（論文審査の結果の要旨）

統計的機械翻訳（SMT）では対訳コーパスから翻訳知識を獲得するため、翻訳の精度は対訳コーパスの量と質に依存する。しかしながら、大規模かつ高品質な対訳コーパスが存在する言語対やドメインは少ない。この問題を解決するために、コンパラブルコーパスを利用することが考えられる。コンパラブルコーパスは各言語独立に、特定の話題について記述された文書対である。コンパラブルコーパス中には単語、単語列（フラグメント）、文の三種類の対訳データが存在する。本論文は、まず対訳単語対を抽出し、これを用いて対訳文・対訳フラグメントを逐次的に抽出するという考え方により、コンパラブルコーパスから対訳データを統合的に抽出するフレームワークを研究し、翻訳の精度を向上させ、その成果をまとめたものである。得られた主要な成果は以下の通りである。

１．対訳単語対抽出において、トピックと文脈知識を用いた反復的抽出手法を提案した。提案手法は種となる事前知識（対訳辞書など）が不要で、抽出の性能が反復的に改善できる。日英、中英、日中のWikipediaデータでの実験により、提案手法の有効性を示した。また、抽出した対訳単語対は後の対訳フラグメントおよび対訳文抽出に使用した。

２．Wikipediaデータから日中対訳コーパスを構築するための堅牢な対訳文抽出システムを提案した。提案システムは対訳文候補のフィルタリングおよび対訳文であるかどうかを識別する分類器から構成されている。システムの性能を向上させるために、日中共通漢字によるフィルタリングと３つの新しい素性セットを提案した。実験では、対訳文抽出の性能と翻訳精度向上の２つの観点から、提案システムの有効性を示した。

３．単語アライメントにより抽出された対訳フラグメント候補を、すでに抽出されている対訳単語対を用いてフィルタリングすることにより、高精度に対訳フラグメントを抽出するシステムを提案した。またカバレッジを向上させるために、日中共通漢字も使用した。日中コンパラブルコーパスで行われた実験の結果、提案システムが対訳フラグメントを正確に抽出し、これを利用することで翻訳の精度も向上することを確認した。また、コンパラブルコーパスにおいては、対訳文と対訳フラグメントが統合的に抽出できることも示した。

よって、本論文は博士（情報学）の学位論文として価値あるものと認める。また、平成２７年２月２７日実施した論文内容とそれに関連した試問の結果、合格と認めた。

注)論文審査の結果の要旨の結句には、学位論文の審査についての認定を明記すること。
更に、試問の結果の要旨（例えば「平成　年　月　日論文内容とそれに関連した口頭試問を行った結果合格と認めた。」）を付け加えること。

Ｗｅｂでの即日公開を希望しない場合は、以下に公開可能とする日付を記入すること。

要旨公開可能日：　　年　　月　　日以降